Explaining the Perfect Sampler

George CASELLA, Michael LAVINE, and Christian P. ROBERT

In 1996, Propp and Wilson introduced coupling from the past (CFTP), an algorithm for generating a sample from the exact stationary distribution of a Markov chain. In 1998, Fill proposed another so–called *perfect sampling* algorithm. These algorithms have enormous potential in Markov Chain Monte Carlo (MCMC) problems because they eliminate the need to monitor convergence and mixing of the chain. This article provides a brief introduction to the algorithms, with an emphasis on understanding rather than technical detail.

KEY WORDS: Coupling from the past; Fill's algorithm; Markov chain Monte Carlo; Stochastic processes.

1. SETTING

A Markov chain is a sequence of random variables $\{X_t\}$ that can be thought of as evolving over time, and where the distribution of X_{t+1} depends on X_t, but not on $X_{t-1}, X_{t-2}, \ldots$. When used in Markov chain Monte Carlo (MCMC) algorithms, Markov chains are usually constructed from a *Markov transition kernel* K, a conditional probability density on a state space $\mathcal{X}$ such that $X_{t+1}|X_t \sim K(X_t, \cdot)$. Interest is usually in the *stationary distribution* of the chain, the distribution π that satisfies

$$\int_{\mathcal{X}} K(x, B)\, d\pi(x) = \pi(B) \quad \text{for any measurable subset } B \text{ of } \mathcal{X}.$$

Thus, if $X_t \sim \pi$, then $X_{t+1} \sim \pi$. In a common application π is the posterior distribution from a Bayesian analysis and K is constructed to have stationary distribution π.

Here is an example that we follow throughout the article.

Example: Beta-Binomial. Following Casella and George (1992), and for some suitable parameters n, α, and β, let $\theta \sim \text{Beta}(\alpha, \beta)$ and $X|\theta \sim \text{Bin}(n, \theta)$, leading to the joint density

$$\pi(x, \theta) \propto \binom{n}{x} \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1},$$

and the conditional density $\theta|x \sim \text{Beta}(\alpha + x, \beta + n - x)$.

We can construct a Markov chain—in fact, a Gibbs sampler—having π as its stationary distribution by using the following transition rule for $(X_t, \theta_t) \mapsto (X_{t+1}, \theta_{t+1})$:

1. choose $\theta_{t+1} \sim \text{Beta}(\alpha + x_t, \beta + n - x_t)$, and
2. choose $X_{t+1} \sim \text{Bin}(n, \theta_{t+1})$.

This transition rule has transition kernel

$$K((x_t, \theta_t), (x_{t+1}, \theta_{t+1})) = f((x_{t+1}, \theta_{t+1})|(x_t, \theta_t)) \propto \binom{n}{x_{t+1}} \theta_{t+1}^{x_{t+1}+\alpha+x_t-1}(1-\theta_{t+1})^{\beta+2n-x_t-x_{t+1}-1}.$$

For future reference we note that the subchain $\ldots, X_t, X_{t+1}, \ldots$ is a Markov chain with $X_{t+1}|x_t \sim \text{BetaBin}(n, \alpha + x_t, \beta + n - x_t)$ and transition kernel

$$K(x_t, x_{t+1}) = f(x_{t+1}|x_t) \propto \binom{n}{x_{t+1}} \times \frac{\Gamma(\alpha+\beta+n)\Gamma(\alpha+x_t+x_{t+1})\Gamma(\beta+2n-x_t-x_{t+1})}{\Gamma(\alpha+x_t)\Gamma(\beta+n-x_t)\Gamma(\alpha+\beta+2n)}.$$

Theorems about stationary distributions and ergodicity apply when the Markov chain satisfies the three properties of *irreducibility*, *reversibility*, and *aperiodicity*, defined in the Appendix. See Robert and Casella (1999, chap. 4) for a brief description or Meyn and Tweedie (1993) and Resnick (1992) among others for book-length treatments. These properties are assumed true for the rest of this article.

The *stationary distribution* of the Markov chain is also a *limiting distribution*: X_t converges in distribution to $X \sim \pi$. For MCMC purposes two useful consequences of our assumptions are that $1/M \sum_{j=1}^{M} h(X_j) \to \text{E}_\pi[h(X)]$ (sometimes called the ergodic theorem) and that a central limit theorem holds.

It is typical in practice to have MCMC algorithms begin from an arbitrarily chosen state at time $t = 0$, say, and run for a long time T, say, in the hope that X_T is a draw approximately from π. One typically discards $X_0, \ldots, X_{T-1}$ and estimates $\text{E}_\pi[h(X)]$ as $1/M \sum_{j=T}^{T+M-1} h(X_j)$. A serious practical problem is determining the "burn-in" time T; see Jones and Hobert (2001). A second practical problem is determining the correlation between X_t and X_{t+1}, which is used to calculate the variance of the estimate. Perfect sampling avoids both problems because it produces independent draws having distribution π precisely.

George Casella is Professor and Chair, Department of Statistics, University of Florida, PO Box 118545, Gainesville, FL 32611-8545 (E-mail: casella@stat.ufl.edu). Michael Lavine is Professor, ISDS, Duke University, Box 90251, Durham, NC (E-mail: michael@stat.duke.edu). Christian P. Robert is Professor, Universite, Paris 9, Dauphine (E-mail: xian@ceremade.dauphine.fr). The first author was supported by National Science Foundation Grant DMS-9625440. We thank the editor and two referees for valuable comments that improved the presentation.

Indeed, the major drawback with using MCMC methods is that their validity is only asymptotic: if we run the sampler kernel until the end of time, we are bound to explore the entire distribution of interest; but, since computing and storage resources are not infinite, we are bound to stop the MCMC sampler at some point. The influence of this stopping time on the distribution of the chain is not harmless and in some cases may induce serious biases (Roberts and Rosenthal 1998). Perfect sampling alleviates this difficulty by producing, in a finite number of steps, exactly the same chain as one running an infinite number of steps, by simply replacing the starting time with $-\infty$ and ∞ with 0. And, at no additional cost, it also removes the dependence on the starting value! In other words, the burn-in time becomes infinite, the chain is in the stationary distribution at time 0, and we produce a sample from the *exact* stationary distribution.

2. COALESCENCE

The first step in obtaining a perfect sample is to find a way to make X_t independent of the starting value. One way to do this is to work with coupled parallel chains.

To illustrate this idea, suppose that the state space $\mathcal{X}$ is finite with k states, and we start a Markov chain in each state at time $t = 0$. These are *parallel chains*. Parallel chains can be *coupled* through a *transition rule* ϕ and random numbers U_t. A transition rule determines X_{t+1} as a function of X_t and U_{t+1}. Note that the same ϕ and same $\ldots, U_t, U_{t+1}, \ldots$ are used for each chain. A common and convenient choice is to let $U_{t+1} \sim \text{Uniform}(0,1)$ and take $X_{t+1} = \phi(x_t, u_{t+1}) = F^{-1}_{X_{t+1}|x_t}(u_{t+1})$, the inverse-cdf function of $X_{t+1}|x_t$ determined by the kernel K and a linear ordering on $\mathcal{X}$. For illustration we return to the Beta-Binomial example.

Example: Beta-binomial, continued. Consider the subchain $\{X_t : t \geq 0\}$ from the previous example, and let $n = 2$, $\alpha = 2$ and $\beta = 4$. The state space is $\mathcal{X} = \{0, 1, 2\}$. The transition probabilities are given by the transition matrix

$$P = \begin{pmatrix} .583 & .333 & .083 \\ .417 & .417 & .167 \\ .278 & .444 & .278 \end{pmatrix},$$

and the cdf matrix

$$C = \begin{pmatrix} .583 & .917 & 1.0 \\ .417 & .833 & 1.0 \\ .278 & .722 & 1.0 \end{pmatrix},$$

in which $p_{ij} = \Pr[X_{t+1} = j - 1 | X_t = i - 1]$ and $c_{ij} = \Pr[X_{t+1} \leq j - 1 | X_t = i - 1]$. The entries of C are the break points at which the behavior of the chain changes. Thus, we can simulate $U_{t+1} \sim \text{Uniform}(0,1)$ and make the transitions illustrated by Figure 1.

Figure 1 shows that coupled chains will all go to the same state, or *coalesce* if there is ever a time t such that either $U_t < .278$, $.583 < U_t < .722$, or $U_t > .917$. Once coupled chains coalesce at time t, they remain coalesced at all times greater than t. And because the U_t's are mutually independent coalescence is guaranteed to happen eventually. Let $X^{s,j} \equiv \{X_t^{s,j}\}_{t \geq s}$ denote a Markov chain that begins in state j at time s. Coalescence is the event that for some t, $X_t^{s,1} = X_t^{s,2} = \cdots = X_t^{s,k}$. The next theorem gives some general results about coalescence.

Theorem 1. Suppose we have k coupled Markov chains, $X^{s,1}, X^{s,2}, \ldots, X^{s,k}$, where

1. $X^{s,j}$ starts in state j at time s (so one chain starts in each state of $\mathcal{X}$);
2. updating is performed according to $X_{t+1}^{s,j} = \phi(X_t^{s,j}, U_{t+1})$, where the U_t's are mutually independent.

Then

(a). The time T to coalescence is a random variable that depends only on $U_1, U_2, \ldots$.
(b). The random variable X_T, the common value at coalescence, is independent of any starting values.

Proof: Part (a) is immediate by construction, and part (b) follows since X_T is a function only of $U_1, \ldots, U_T$ and not of the j in $X_s^{s,j}$.

Conclusion (b) of Theorem 1 says that T is a time at which the initial state of the chain has "worn off." One might therefore hope that X_T is a draw from the stationary distribution π. This hope is false. It is true that if T^* is a *fixed* time, and X_{T^*} is independent of $X_s^{s,j}$, then $X_{T^*} \sim \pi$. Unfortunately, T is a random time and in general, $X_T \not\sim \pi$, as the following example illustrates.

Example: Two-state. Consider the Markov chain with state space $\{1, 2\}$ and transition kernel $K(1,1) = K(1,2) = .5$; $K(2,1) = 1$; $K(2,2) = 0$. The stationary distribution is $\pi(1) = 2/3$; $\pi(2) = 1/3$. A little thought shows that parallel chains can coalesce only in $X_T = 1$ and therefore $X_T \not\sim \pi$.

3. PROPP AND WILSON

Propp and Wilson (1996) explained how to take advantage of coalescence while sampling the chain at a fixed time, thereby producing a random variable having distribution π, exactly. Their algorithm is called *coupling from the past* (CFTP), and is based on the idea that if a chain were started at time $t = -\infty$ in any state $X_{-\infty}$, it would be in equilibrium by time $t = 0$, so X_0 would be a draw from π. This would happen since the chain would have run for an infinite length of time.

To implement this idea in an algorithm, we use the coalescence strategy. We first find a time $-T$ such that $X_0 \equiv X_0^{-T,j}$ does not depend on $X_{-T}^{-T,j}$ (coalescence occurs between time $-T$ and time 0), and then we determine X_0 by starting chains from all states at time $t = -T$ and following them to time $t = 0$.

CFTP is an algorithm for finding $-T$ and X_0 and goes as follows.

1. Start chains $X^{-1,1}, X^{-1,2}, \ldots, X^{-1,k}$ at time $t = -1$ from every state of $\mathcal{X}$. Generate U_0.

2. Update each chain to time $t = 0$ by applying the transition rule $X_0^{-1,j} = \phi(X_{-1}^{-1,j}, U_0)$. If the chains have coalesced at time $t = 0$, then $T = -1$ and the common value X_0 is a draw from π.

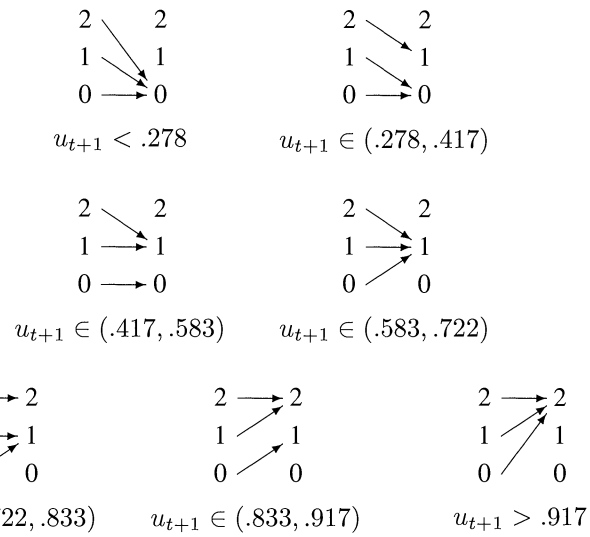


Figure 1. All possible transitions for the Beta-Binomial(2,2,4) example.

3. Otherwise, move back to time $t = -2$, start chains $X^{-2,1}, \ldots, X^{-2,k}$, generate U_{-1}, and update each chain using $X_{-1}^{-2,j} = \phi(X_{-2}^{-2,j}, U_{-1})$ and $X_0^{-2,j} = \phi(X_{-1}^{-2,j}, U_0)$. If the chains have coalesced at time $t = 0$, then $T = -2$ and the common value X_0 is a draw from π.

4. Otherwise, move back to time $t = -3$ and continue.

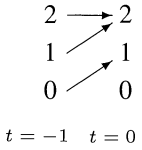
It is crucial, in Step 3, to use the same U_0 from Step 1. Specifically, we start chains at time $t = -2$ from every state; draw U_{-1}; use U_{-1} to update all the chains to time $t = -1$; use the U_0 from step (1) to update all the chains to time $t = 0$; check for coalescence; and either accept $T = -2$ and X_0 if the chains have coalesced or go back to time $t = -3$ if they have not. The algorithm continues backing through time until coalescence occurs.

Theorem 2. The CFTP algorithm returns a random variable distributed exactly according to the stationary distribution of the Markov chain.

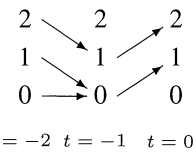
Proof: See the Appendix.

We use the Beta-Binomial example for illustration.

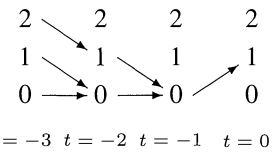
Example: Beta-Binomial, continued. Begin at time $t = -1$ and draw U_0. Suppose $U_0 \in (.833, .917)$. The next picture shows the result of updating all chains.



The chains have not coalesced, so we go to time $t = -2$ and draw U_{-1}. Suppose $U_{-1} \in (.278, 417)$. The next picture shows the result of updating all chains.



The chains have still not coalesced so we go to time $t = -3$. Suppose $U_{-2} \in (.278, .417)$. The next picture shows the result of updating all chains.



All chains have coalesced into $X_0 = 1$. We accept X_0 as a draw from π. Note that even though the chains have coalesced at $t = -1$, we do not accept $X_{-1} = 0$ as a draw from π.

In CFTP, T and X_0 are dependent random variables. Therefore, a user who gets impatient or whose computer crashes and who therefore restarts runs when T gets too large will generate biased samples. Another algorithm, due to Fill (1998), generates samples from π in a way that is independent of the number of steps.

4. FILL'S ALGORITHM

A simple version of Fill's algorithm (Fill) is:

1. Arbitrarily choose a time $T > 0$ and state $x_T = z$.
2. Generate $X_{T-1}|x_T$, $X_{T-2}|x_{T-1}, \ldots, X_0|x_1$.
3. Generate $[U_1|x_0, x_1]$, $[U_2|x_1, x_2], \ldots, [U_T|x_{T-1}, x_T]$.
4. Begin chains $X^{0,1}, \ldots, X^{0,k}$ in all states at time 0 and use the common $U_1, \ldots, U_T$ to update all chains.
5. If the chains have coalesced by time T (and are in state z at time T), then accept x_0 as a draw from π.
6. Otherwise begin again, possibly with a new T and z.

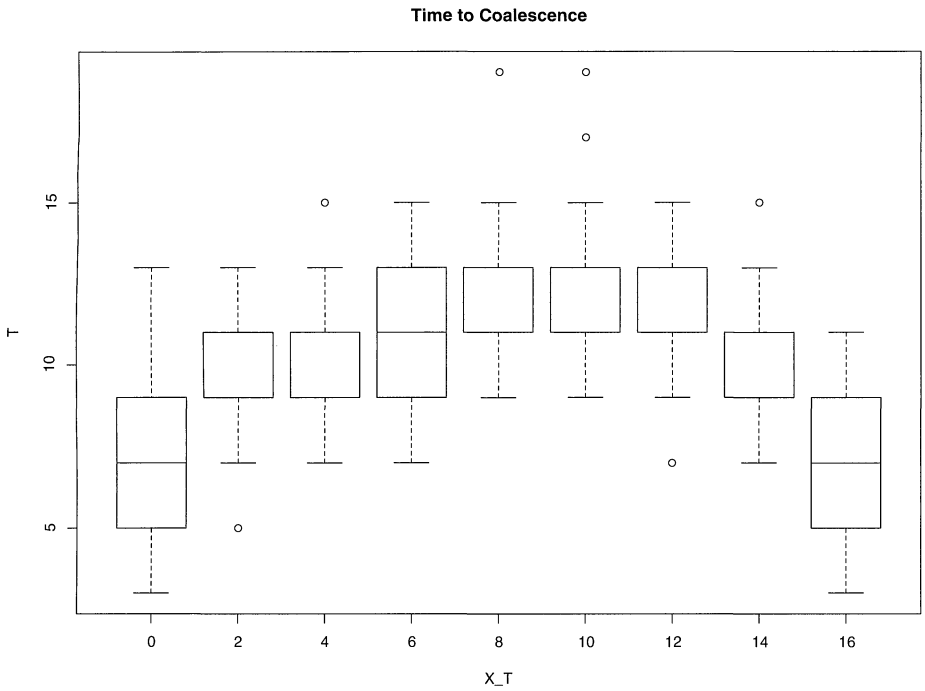


Figure 2. Time to coalescence for 50 runs of Fill's algorithm, for each value of X_T.

We note that the $U_1, \ldots, U_T$ used for the coalescing chains are generated in such a way to ensure that $x \to z$; that is, $X_T^{0,x} = z$. So, for example, generate U_1 to be uniform on the set $\{u : x_1 = \phi(x_0, u)\}$, U_2 to be uniform on the set $\{u : x_2 = \phi(x_1, u)\}$ and so on. See the example for a further illustration.

There are two ways to prove that Fill is correct. We present one proof here and the second proof in the appendix. Let $C_T(z)$ be the event that all chains have coalesced and are in state z at time T; that is, $X_T^{0,j} = z$ for all j.

First proof: Fill delivers a value only if it accepts $X_0 = x$, so we need to prove $\Pr[X_0 = x|\text{accept}] = \pi(x)$. This probability is

$$\Pr[X_0 = x|\text{accept}] = \frac{\Pr[z \to x]\Pr[C_T(z)|x \to z]}{\sum_{x'} \Pr[z \to x']\Pr[C_T(z)|x' \to z]}.$$

Now because the coalescence event entails each $x' \to z$, we have for every x'

$$\Pr[C_T(z)|x' \to z] = \frac{\Pr[C_T(z) \text{ and } x' \to z]}{\Pr[x' \to z]} = \frac{\Pr[C_T(z)]}{\Pr[x' \to z]}, \quad (1)$$

and writing $\Pr[x' \to z] = K^T(x', z)$ the probability becomes

$$\begin{aligned}\Pr[X_0 = x|\text{accept}] &= \frac{K^T(z,x)\Pr[C_T(z)]/K^T(x,z)}{\sum_{x'} K^T(z,x')\Pr[C_T(z)]/K^T(x',z)} \\ &= \frac{K^T(z,x)/K^T(x,z)}{\sum_{x'} K^T(z,x')/K^T(x',z)},\end{aligned}$$

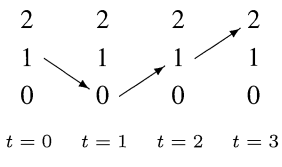
From the detailed balance condition (see the Appendix) $\pi(z)K^T(z,x) = \pi(x)K^T(x,z)$ we have $K^T(z,x)/K^T(x,z) = \pi(x)/\pi(z)$, and thus,

$$\Pr[X_0 = x|\text{accept}] = \frac{\pi(x)/\pi(z)}{\sum_{x'} \pi(x')/\pi(z)} = \pi(x).$$

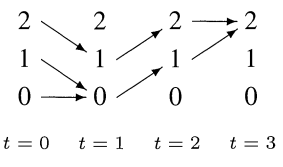
We follow the Beta-binomial (2,2,4) example through the steps in Fill.

Example: Beta-Binomial, continued.

1. We arbitrarily choose $T = 3$ and $X_T = 2$.
2. Our chain is reversible, so $[X_2|X_3 = 2] = [X_3|X_2 = 2] = \text{BetaBin}(2,4,4)$. The probabilities are given on page 300. We generate X_2. Suppose it turns out to equal 1. Similarly, $X_1|X_2 = 1 \sim \text{BetaBin}(2,3,5)$; suppose we get $X_1 = 0$; $X_0|X_1 = 2 \sim \text{BetaBin}(2,4,4)$; suppose we get $X_0 = 1$. The next picture shows the transitions we have generated.



3. $X_0 = 1$, $X_1 = 0$, $X_2 = 1$ and $X_3 = 2$ imply $U_1 \sim \text{U}(0, .417)$; $U_2 \sim \text{U}(.583, .917)$; and $U_3 \sim \text{U}(.833, 1)$. (See Figure 1.) Suppose we generate $U_1 \in (.278, .417)$, $U_2 \in (.833, .917)$ and $U_3 > .917$.
4. Begin chains in states 0, 1, and 2.
5. The next picture follows the chains through time $t = 3$.



6. The chains coalesce in $X_3 = 2$; so we accept $X_0 = 1$ as a draw from π.

Fill depends on an arbitrary choice of T and X_T. To get some feeling for how big T needs to be and whether the choice of X_T is important, we ran Fill on a Beta-binomial$(16, 2, 4)$ example. For each of $X_T = 0, 2, \ldots, 16$, we ran Fill in a loop with $T = 1, 3, \ldots$ successively until the algorithm returned a value. The whole simulation was repeated 50 times. Figure 2 is a boxplot, sorted by X_T, of the T for which coalescence was achieved. The horizontal axis is the value of X_T which we fixed in advance. The vertical axis is the value of T for which coalescence occured. The figure shows that coalescence occured much more quickly when we chose either $X_T = 0$ or $X_T = 16$ than any other value of X_T.

5. DISCUSSION

- A potentially troublesome point is detecting whether coalescence has occurred. In general, starting and keeping track of chains from every state is computationally infeasible. In (partially) ordered state spaces with a *monotone* transition rule it is only necessary to keep track of chains started from the maximal and minimal members. A monotone transition rule is one in which $X_t \geq Y_t \Rightarrow X_{t+1} = \phi(X_t, u_{t+1}) \geq Y_{t+1} = \phi(Y_t, u_{t+1})$. If we use an inverse-cdf function ϕ (with an appropriate linear order) and the kernel K is stochastically monotone, then the transition rule will be monotone.

This is the case in our example, where a chain started from state 1 is sandwiched between chains started from states 0 and 2. Therefore it is only necessary to keep track of chains started from 0 and 2 to determine whether coalescence has occurred. In fact, if there exist maximal and minimal elements, coalescence is detectable even with a continuous state space. Nonmonotone transition rules or state spaces without minimal and maximal elements require more sophisticated methods. See Fill et al. (1999) or Green and Murdoch (1999) for details and extensions.

- In describing CFTP we set T successively equal to $-1, -2, \ldots$. In fact, any decreasing sequence would do as well. Propp and Wilson (1996) argued that $T = -1, -2, -4, -8, \ldots$ is near optimal. In Fill, if X_0 is rejected, or if many realizations are needed, it may be better to choose new values of T and z for the next proposal. Figure 2 shows that some combinations of (T, z) are more likely to lead to coalescence than others. There is no general theory at present to guide the choice of (T, z). In practice the results of early iterations may guide the choice of (T, z) in later iterations.

- In his original algorithm described here, when running the k chains for coalescence, Fill used constrained uniform variables $U_1, \ldots, U_T$ conditional on $X_0, \ldots, X_T$, generating $[U_1|x_0, x_1]$, $[U_2|x_1, x_2]$, $\ldots$, $[U_T|x_{T-1}, x_T]$. This ensures that the chain starting in x will end up in z. This is practical as long as it is not too difficult to sample from the conditional distribution of the U_i's given the X_i's.

An alternative to the algorithm described in Fill is to generate the U_i's unconditionally. (Typically $U_i \sim U(0, 1)$.) Using these U_i's, check whether $x_0 \to z$. If yes, then also check for $C_T(z)$ and either accept or reject X_0 accordingly. Otherwise, discard the U_i's and generate another set until finding one such that $x_0 \to z$. Ultimately we will accept x_0 with probability $\Pr[C_T(z)|x_0 \to z]$, as required. However, the implementation of such an alternative is typically impractical in real applications.

- Some practical applications of Markov chains iterate between a discrete X and a parameter θ that might be either discrete or continuous. In such cases we can obtain perfect samples from the joint distribution of both X and θ. In our Beta-Binomial example, once we have a perfect sample of X we can obtain a perfect sample of θ by sampling from $[\theta|X]$. For a more interesting example, consider modeling the data Y as a mixture of Normal distributions. The model is usually extended to include indicator variables X, which are not observed but which indicate which Y's come from the same mixture components. Conditional on X, the model is a straightforward collection of Normals. Let θ denote all unknown parameters other than X. The posterior is typically analyzed through a Gibbs sampler that iterates between $[X|\theta]$ and $[\theta|X]$. The iterates of X form a subchain on a finite state space and are amenable to perfect sampling. Given a perfect sample of X, one can simulate from $[\theta|X]$ to obtain a perfect sample of θ.

This remark extends to other latent variable models, but one must keep in mind that the size of the finite parameter space of X in the mixture example is k^n, which rapidly gets unmanageable unless monotonicity features can be exhibited, as in Hobert, Robert, and Titterington (1999).

- To remove the difficulty with continuous state space chains, another promising direction relies on *slice sampling*. This technique is a special case of Gibbs sampling (see Robert and Casella 1999, sect. 7.1.2) and takes advantage of the fact that the marginal (in X) of the uniform distribution on $\{(x, u);\ u \leq \pi(x)\}$ is $\pi(x)$. The idea, detailed by Mira, Møller, and Roberts (1999), is that, if X_0' is a variable generated from the uniform distribution on $\{x;\ \pi(x) \geq \epsilon\pi(x_0)\}$, it can also be taken as a variable generated from the uniform distribution on $\{x;\ \pi(x) \geq \epsilon\pi(x_1)\}$ for all x_1's such that $\epsilon\pi(x_0) \leq \epsilon\pi(x_1) \leq \pi(x_0')$ by a simple accept–reject argument. Therefore, assuming a bounded state space $\mathcal{X}$, if one starts with X_0' generated uniformly on $\mathcal{X}$, a finite sequence $X_0', \ldots, X_T'$ can be used instead of the continuum of possible starting values, with x_i' being generated from a uniform distribution on $\{x;\ \pi(x) \geq \pi(x_{i-1}')\}$, and T being such that $\pi(x_T') \geq \epsilon \sup \pi(x)$. Moreover, slice sampling exhibits natural monotonicity structures which can be exploited to further reduce the number of chains. The practical difficulty of this approach is that uniform distributions on $\{x;\ \pi(x) \geq \epsilon\pi(x_0)\}$ may be hard to simulate, as shown by Casella, Mengersen, Robert, and Titterington (1999) in the setup of mixtures.

- Perfect sampling is currently an active area of research. David Wilson maintains a Web site of papers on perfect sampling at http://dimacs.rutgers.edu:80/~dbwilson/exact.html. The interested reader can find links to articles ranging from introductory to the latest research.

APPENDIX

A.1 A Markov Chain Glossary

We will work with discrete state space Markov chains. The following definitions can be extended to continuous state spaces

as long as the usual measurability complications are carefully dealt with.

A Markov chain $X_1, X_2, \ldots$, is *irreducible* if the chain can move freely throughout the state space; that is, for any two states x and x', there exists an n such that $\Pr[X_n = x' | X_0 = x] > 0$. Moreover, as the chains we are considering are all *positive*, that is, the stationary distribution is a probability distribution, irreducibility also implies that the chain is *recurrent*. A recurrent chain is one in which the average number of visits to an arbitrary state is infinite.

A state x has *period* d if $P(X_{n+t} = x | X_t = x) = 0$ if n is not divisible by d, d being the largest integer with this property. For example, if a chain starts ($t = 0$) in a state with period 3, the chain can only return to that state at times $t = 3, 6, 9, \ldots$. If a state has period $d = 1$, it is *aperiodic*. In an irreducible Markov chain, all states have the same period. If that period is $d = 1$, the Markov chain is aperiodic.

We then have the following theorems.

Theorem A.1: Convergence to the stationary distribution. If the countable state space Markov chain $X_1, X_2, \ldots$, is positive, recurrent, and aperiodic with stationary distribution π, then from every initial state

$$X_n \to X \sim \pi.$$

A positive, recurrent and aperiodic Markov chain is often called *ergodic*, a name also given to the following theorem, a cousin of the Law of Large Numbers.

Theorem A.2: Convergence of sums. If the countable state space Markov chain $X_1, X_2, \ldots$, is ergodic with stationary distribution π, then from every initial state

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \to E_\pi h(X)$$

provided $E_\pi |h(X)| < \infty$.

Adding the property of *reversibility* will get us a central limit theorem. If we reverse the direction of time for a stationary, ergodic Markov chain, the reversed process is also a Markov chain (see Ross 1985, sec. 4.7), but does not necessarily have the same transition probabilities. The reversed chain does have the same transition probabilities and is said to be *reversible* if

$$\pi(y) K(y, x) = \pi(x) K(x, y) \quad \text{for all} \quad x, y.$$

This condition is also known as *detailed balance*, and insures that the transition probabilities are the same whether we go forward or backward along the chain.

Theorem A.3: Central limit theorem. If the countable state space Markov chain $X_1, X_2, \ldots$, is ergodic and reversible with stationary distribution π, then from every initial state

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(X_i) - E_\pi h(X)] \to \mathcal{N}(0, \sigma^2),$$

provided $0 < \sigma^2 = \operatorname{Var} h(X_0) + \sum_{i=1}^{\infty} \operatorname{Cov}_\pi(h(X_0), h(X_i)) < \infty$.

A.2 Proof of Theorem 2

Let $\mathcal{L}(X)$ denote the probability law of the random variable X and $\mathcal{L}(X) \to \pi$ denote convergence in probability. The proof is based on establishing the following three facts:

1. The CFTP algorithm will finish in finite time and produce a value; call it X_0.
2. For each j, $\mathcal{L}(X_0^{-t,j}) \to \pi$ as $t \to \infty$.
3. For each j, $X_0^{-t,j} \to X_0$ as $t \to \infty$.

It then follows that $X_0 \sim \pi$.

Fact 1. We adapt the proof presented by Thönnes (1999). By irreducibility, for each j we can find N_j such that

$$P(X_0^{N_j, j} = x) > 0, \quad \text{for all} \quad x \in \mathcal{X}.$$

Set $N = \max\{N_1, N_2, \ldots, N_k\}$. It then follows that each $X_0^{-N,j}$ has positive probability of being in any state, and that for some $\varepsilon > 0$

$$P(X_0^{-N,1} = X_0^{-N,2} = \cdots = X_0^{-N,k}) > \varepsilon.$$

Now run the CFTP algorithm in blocks of size N as follows.

(i). Starting at time $-N$, run the k coupled chains to time 0. If they have not coalesced

(ii). Starting at time $-2N$, run the k coupled chains to time 0. If they have not coalesced.

Define C_i to be the event

$$X_{-(i-1)N}^{-iN,1} = X_{-(i-1)N}^{-iN,2} = \cdots = X_{-(i-1)N}^{-iN,k},$$

that is, the event that k parallel chains started at $t = -iN$ will have coalesced by $t = -(i-1)N$. From the preceding argument we have that $P(C_i) > \varepsilon$. Moreover, the C_i are independent because coalescence in $(-iN, -(i-1)N)$ only depends on $U_{-iN}, U_{-iN-1}, \ldots, U_{-(i-1)N}$ (which are independent of all of the other U_i's).

Finally, we observe that

$$\begin{aligned} P(\text{ no coalescence after } I \text{ iterations}) &\le \prod_{i=1}^{I} [1 - P(C_i)] \\ &< (1 - \varepsilon)^I \\ &\to 0 \text{ as } I \to \infty, \end{aligned}$$

showing that the probability of coalescence is 1. We can, in fact, draw the stronger conclusion that the coalescence time is almost surely finite by noting that

$$\sum_{i=1}^{\infty} P(C_i) = \infty \Rightarrow P(C_i \text{ infinitely often}) = 1,$$

from the Borel–Cantelli Lemma.

Fact 2. We next show that for $j = 1, 2, \ldots, k$,

$$\mathcal{L}(X_0^{-t,j}) \to \pi \quad \text{as} \quad t \to \infty.$$

But $\mathcal{L}(X_0^{-t,j}) = \mathcal{L}(X_t^{0,j})$ because they are both the distribution of a Markov chain that starts in state j and progresses through t time steps. And $\mathcal{L}(X_t^{0,j}) \to \pi$ because π is the stationary distribution.

Fact 3. Fact 1 says there exists an N such that coalescence occurs between time $-N$ and time 0. Therefore, for all $t \geq N$, $X_0^{-t,j} = X_0$, which implies Fact 3.

A.3 Alternate Proof of Fill

We can view Fill as a rejection algorithm: generate and propose $X_0 = x$; then accept x as a draw from π if $C_T(z)$ has occurred. The proposal distribution is the T-step transition density $K^T(z, \cdot)$. Fill is a valid rejection algorithm if we accept $X_0 = x$ with probability

$$\frac{1}{M}\frac{\pi(x)}{K^T(z,x)} \quad \text{where} \quad M \geq \sup_x \frac{\pi(x)}{K^T(z,x)}.$$

From detailed balance we can write $\pi(x)/K^T(z,x) = \pi(z)/K^T(x,z)$ and, since $\Pr[C_T(z)] \leq K^T(x',z)$ for any x' we have the bound

$$\frac{\pi(x)}{K^T(z,x)} = \frac{\pi(z)}{K^T(x,z)} \leq \frac{\pi(z)}{\Pr[C_T(z)]} \equiv M.$$

So we accept $X_0 = x$ with probability $\frac{1}{M}\frac{\pi(x)}{K^T(z,x)}$, which is quite difficult to compute. However,

$$\begin{aligned}\frac{1}{M}\frac{\pi(x)}{K^T(z,x)} &= \frac{\Pr[C_T(z)]}{\pi(z)}\frac{\pi(x)}{K^T(z,x)} \\ &= \frac{\Pr[C_T(z)]}{\pi(z)}\frac{\pi(z)}{K^T(x,z)} = \frac{\Pr[C_T(z)]}{K^T(x,z)},\end{aligned}$$

where we have again used detailed balance. But now, from (1), we have that $\frac{\Pr[C_T(z)]}{K^T(x,z)} = \Pr[C_T(z) | x \to z]$, exactly the event that Fill simulates.

Finally, note that the algorithm is more efficient if the acceptance probability $1/M$ is as large as possible, so choosing z to be the state that maximizes $\Pr[C_T(z)]/\pi(z)$ is a good choice. This, also, will be a difficult calculation, but in running the algorithm, these probabilities can be estimated.

[Received June 2000. Revised March 2001.]

REFERENCES

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46 167–174.

Casella, G., Mengersen, K. L., Robert, C. P., and Titterington, D. M. (1999), "Perfect Sampling for Mixtures," Technical Report, CREST, Insee, 1999.

Fill, J. A. (1998), "An Interruptible Algorithm for Perfect Sampling via Markov Chains," *Annals of Applied Probability*, 8, 131–162.

Fill, J. A., Machida, M., Murdoch, D. J., and Rosenthal, J. S. (1999), "Extension of Fill's Perfect Rejection Sampling Algorithm to General Chains," Technical Report, The Johns Hopkins University, Dept. of Mathematical Sciences.

Green, P. J., and Murdoch, D. J. (1999), "Exact Sampling for Bayesian Inference: Towards General Purpose Algorithms," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Clarendon Press.

Hobert, J. P., Robert, C. P., and Titterington, D. M. (1999), "On Perfect Simulation for Some Mixtures of Distributions," *Statistics and Computing*, 9, 287–298.

Jones, G., and Hobert, J. (2001), "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo," Technical Report, Department of Statistics, Univeristy of Florida.

Meyn, S. P., and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, New York: Springer-Verlag.

Mira, A., Møller, J., and Roberts, G. O. (1999), "Perfect Slice Sampler," Technical Report R-99-2020, Dept. of Mathematical Science, Aalborg University.

Propp, J. G., and Wilson, D. B. (1996), "Exact Sampling With Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 223–252.

Resnick, S. I. (1992), *Adventures in Stochastic Processes*, Boston: Birkhäuser.

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Roberts, G. O., and Rosenthal, J. S. (1998), "Markov Chain Monte Carlo: Some Practical Implications of Theoretical Results" (with discussion), *Canadian Journal of Statistics*, 26, 5–32.

Ross, S. M. (1985), *Probability Models* (3rd ed.), New York: Academic Press.

Thönnes, E. (1999), "A Primer on Perfect Sampling," Technical Report, Department of Mathematical Statistics, Chalmers University of Technology.